Journal of Complexity 25 (2009) 188–200

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

Learning from uniformly ergodic Markov chains☆

Bin Zou[a,b,*], Hai Zhang[a], Zongben Xu[a]

[a] Institute for Information and System Science, Faculty of Science, Xi'an Jiaotong University, Xi'an, 710049, PR China
[b] Faculty of Mathematics and Computer Science, Hubei University, Wuhan, 430062, PR China

ARTICLE INFO

Article history:
Received 16 April 2008
Accepted 13 January 2009
Available online 30 January 2009

Keywords:
ERM algorithms
Uniform ergodic Markov chain samples
Generalization bound
Uniform convergence
Relative uniform convergence

ABSTRACT

Evaluation for generalization performance of learning algorithms has been the main thread of machine learning theoretical research. The previous bounds describing the generalization performance of the empirical risk minimization (ERM) algorithm are usually established based on independent and identically distributed (i.i.d.) samples. In this paper we go far beyond this classical framework by establishing the generalization bounds of the ERM algorithm with uniformly ergodic Markov chain (u.e.M.c.) samples. We prove the bounds on the rate of uniform convergence/relative uniform convergence of the ERM algorithm with u.e.M.c. samples, and show that the ERM algorithm with u.e.M.c. samples is consistent. The established theory underlies application of ERM type of learning algorithms.

1. Introduction

Recently Support Vector Machines (SVMs) have become one of the most widely used algorithms in the machine learning community [1]. Besides their good performance in practical applications they also enjoy a good theoretical justification in terms of both universal consistency and learning rates when training samples come from an i.i.d. process. This renewed interest for theory naturally boosted the development of performance bounds for learning algorithms (see [2–6], etc.). However, this i.i.d. assumption cannot often be strictly justified in real-world applications, and many machine learning applications such as market prediction, system diagnosis, and speech recognition are inherently temporal in nature, and consequently not i.i.d. processes [7]. Relaxations of such i.i.d. assumption have

☆ Supported by National 973 project (2007CB311002), NSFC key project (70501030) and Foundation of Hubei Educational Committee (Q200710001).

* Corresponding author at: Faculty of Mathematics and Computer Science, Hubei University, Wuhan, 430062, PR China.
E-mail addresses: zoubin0502@hubu.edu.cn (B. Zou), zhanghai@nwu.edu.cn (H. Zhang), zbxu@mail.xjtu.edu.cn (Z. Xu).

0885-064X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2009.01.001

been considered for quite a while in both machine learning and statistics literatures. For example, Yu [8] established the rates of convergence for empirical processes of stationary mixing sequences. Modha and Masry [9] established the minimum complexity regression estimation with m-dependent observations and strongly mixing observations respectively. Vidyasagar [10] considered the notions of mixing and proved that most of the desirable properties (e.g. PAC or UCEMUP property) of i.i.d. sequence are preserved when the underlying sequence is mixing sequence. Steinwart, Hush and Scovel [7] proved that the SVMs for both classification and regression are consistent only if the data-generating process satisfies a certain type of law of large numbers (e.g. WLLNE, SLLNE). Smale and Zhou [11] considered online regularization learning algorithm based on Markov sampling. Zou and Li [12] established the bounds on the rate of uniform convergence of learning machines with strongly mixing observations. Zou, Li and Xu [13] obtained the generalization bounds of the ERM algorithm with exponentially strongly mixing observations.

There have been many dependent (not i.i.d.) sampling mechanisms studied in machine learning literatures ([14,15] etc.). In the present paper we focus on an analysis in the case when the samples are Markov chains (that is, the Markov chain samples). The Markov chain samples appear so often and naturally in applications, especially in biological (DNA or protein) sequence analysis, speech recognition, character recognition, content-based web search and marking prediction. Two examples are as follows:

Example 1. Consider the problem of an insurance company wanting to draft the amount of insurance money and claim settlement according to the health condition of insurance applicants. In the simplest case, the health condition of an insurance applicant consists of healthy and ill. For an insurance applicant during given age stage, we suppose that the probability that he/she is healthy this year and also next year is given. The probability that he/she is ill this year but healthy next year is also known. Let x_i be the health condition given by the ith year, and y_i be the corresponding profit or loss the insurance company made. Then $\{x_i\}$ is a sequence with Markov property. The insurance company had a data set of past insurance applicants and the profit or loss of the company. To draft the amount of insurance money and claim settlement, one should learn the unknown functional dependency between x_i and y_i from the Markov chain samples $\{\mathbf{z}_i = (x_i, y_i)\}_{i\geq 1}$.

Example 2. We usually have the following quantitative example in the models of random walk and predicting the weather, that is, suppose that $\{x_i\}$ is a Markov chain consisting of five states 1, 2, 3, 4, 5 and having transition probability matrix

$$P = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.1 & 0.3 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.1 & 0.3 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.1 & 0.3 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.1 & 0.3 \end{bmatrix}.$$

By the matrix P and Matlab software, we can create a sequence with Markov property, for example, $x_1 = 1, x_2 = 1, x_3 = 5, x_4 = 3, \ldots$. Through target function $y = f(x) = x^2 + 10x + 3$, we also can produce the corresponding values of x_i, that is, $y_1 = 14, y_2 = 14, y_3 = 78, y_4 = 42, \ldots$. Then a problem is posed: how can we learn the target function $f(x) = x^2 + 10x + 3$ from the Markov chain samples

$$S = \{\mathbf{z}_1 = (1, 14), \mathbf{z}_2 = (1, 14), \mathbf{z}_3 = (5, 78), \mathbf{z}_4 = (3, 42), \ldots\}.$$

Many empirical evidences show that a learning algorithm very often performs well with Markov chain samples ([16,17], etc.). Why it is so, however, has been unknown (particularly, it is unknown how well it performs in terms of consistency and generalization). Answering those questions is the purpose of the present paper. We will provide theoretical justification of the success of the ERM algorithm by establishing a consistency and the generalization bound estimation results of the ERM algorithm with u.e.M.c. samples. Following this schedule, in Section 2 we introduce some notions and notations. In Sections 3 and 4 we derive the bounds on the rate of uniform convergence and relative uniform convergence of the ERM algorithm respectively, and obtain the generalization bounds of the

ERM algorithm with u.e.M.c. samples. Finally, we conclude the paper with some useful remarks in Section 5.

2. Preliminaries

In this section we introduce the definitions and notations used throughout the paper, and present Hoeffding's inequality on u.e.M.c samples, which will be used in the next studies.

Suppose $(\mathbf{Z}, \mathcal{S})$ is a measurable space, a Markov chain is a sequence of random variables $\{\mathbf{z}_t\}_{t\geq 0}$ together with a set of probability measures $P^n(z, A)$, $z \in \mathbf{Z}$, $A \in \mathcal{S}$. It is assumed that

$$P^n(z, A) = \text{Prob}\{\mathbf{z}_{n+i} \in A | \mathbf{z}_j, j < i, \mathbf{z}_i = z\}.$$

Thus $P^n(z, A)$ denotes the probability that the state $\mathbf{z}$ will belong to the set A after n time steps, starting from the initial state z at time i. The fact that the transition probability does not depend on the values of $\mathbf{z}$ prior to time i is the Markov property, that is

$$\text{Prob}\{\mathbf{z}_{n+i} \in A | \mathbf{z}_j, j < i, \mathbf{z}_i = z\} = \text{Prob}\{\mathbf{z}_{n+i} \in A | \mathbf{z}_i = z\},$$

and the fact that the transition probability does not depend on the initial time i means that the Markov chain is stationary [10].

Given two probability measures μ_1, μ_2 on the measurable space $(\mathbf{Z}, \mathcal{S})$, we define the total variation distance between the two measures as follows:

$$\|\mu_1 - \mu_2\|_{TV} = 2 \sup_B |\mu_1(B) - \mu_2(B)|.$$

With these notations, there are several definitions of Markov chain, but we shall be concerned with only one, namely, uniformly ergodic Markov chain in this paper [18,19].

Definition 1 (*[18]*). A Markov chain $\underline{Z} = \{\mathbf{z}_i\}_{i\geq 0}$ is called a uniformly ergodic Markov chain, if

$$\sup_{z\in\mathbf{Z}} \|P^n(z, A) - \pi(A)\|_{TV} \doteq d(n) \to 0 \quad \text{as } n \to \infty$$

for $n \in \mathbb{N}$, and every measurable set $A \in \mathcal{S}$, where π is the stationary distribution of the Markov chain $\underline{Z}$.

Meyn and Tweedie (see Theorem 16.0.2 in [16]), Aldous, Lorász and Winkler (see Theorem B in [19]) proved that Definition 1 is equivalent to the following assumption.

Assumption 1 (*[20]*). There exists a probability measure ψ on $\mathcal{S}$, a positive real number $\lambda \in (0, 1)$, and an integer $m \geq 1$ such that

$$P^m(z, A) \geq \lambda\psi(A)$$

for every $z \in \mathbf{Z}$ and any measurable set $A \in \mathcal{S}$.

Remark 1. Assumption 1 listed here is closely related to the assumption of uniform ergodicity introduced in [16]. Meyn and Tweedie [16] proved that a chain satisfying Assumption 1 automatically possess a unique stationary distribution. To emphasize the role of parameters λ and m, a Markov chain satisfying Assumption 1 is denoted as $\underline{Z}(\lambda, m)$ in what follows.

Given n samples

$$S = \{\mathbf{z}_0 = (x_0, y_0), \mathbf{z}_1 = (x_1, y_1), \ldots, \mathbf{z}_{n-1} = (x_{n-1}, y_{n-1})\} \in \mathbf{Z}^n$$

drawn from the first n samples of the u.e.M.c. $\underline{Z}(\lambda, m)$ according to π, which is a fixed but unknown stationary distribution on $\mathbf{Z} = X \times Y$. The goal of learning from the Markov chain samples S is to find a function f that assigns values to unlabeled samples such that when a new unlabeled sample is given, the function f can forecast it correctly. Let

$$\mathcal{E}(f) = \text{E}[\ell(f, \mathbf{z})] = \int \ell(f, \mathbf{z}) \text{d}\pi$$

be the expected risk (or expected error) of function f, where $\ell(f, \mathbf{z})$, is a non-negative loss function. In machine learning [7], the margin-based loss functions such as the (squared) hinge loss, the AdaBoost loss, the logistic loss and the least square loss are very often used in classification applications, and the distance-based loss functions such as the least squares loss, Huber's insensitive loss, the logistic loss, and the ε-insensitive loss are frequently adopted in regression applications. Because our purpose in the present research is to discuss general learning problems, we consider the loss function of general form $\ell(f, \mathbf{z})$ in the following. Therefore, we now move the focus from a function f to a family $\mathcal{F}$ of such functions.

A learning problem can be formulated as finding the minimizer of the expected risk over a hypothesis space $\mathcal{F}$. Since the distribution π is unknown and we only know the set S of random samples, the minimizer of the expected risk $\mathcal{E}(f)$ cannot be computed directly. The Empirical Risk Minimization (ERM) principle [1] then advocates that instead of minimizing the expected risk, an approximate solution is found through minimizing the so-called empirical risk (or empirical error) defined by

$$\mathcal{E}_n(f) = \frac{1}{n}\sum_{i=0}^{n-1} \ell(f, z_i).$$

Let $\tilde{f}$ be a function minimizing the expected risk $\mathcal{E}(f)$ over $\mathcal{F}$, that is,

$$\tilde{f} = \arg\min_{f\in\mathcal{F}} \mathcal{E}(f) = \arg\min_{f\in\mathcal{F}} \int \ell(f, z)d\pi.$$

We define $\hat{f}$ to be a function minimizing the empirical risk $\mathcal{E}_n(f)$ over $\mathcal{F}$, i.e.,

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \mathcal{E}_n(f) = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=0}^{n-1} \ell(f, z_i). \tag{1}$$

According to the ERM principle, we then consider the function $\hat{f}$ as an approximation of the target function $\tilde{f}$. Thus a central question of the ERM learning (1) is how well $\hat{f}$ really approximate $\tilde{f}$. If this approximation is good, then the ERM algorithm is said to be generalize well. An ERM algorithm with generalization capability implies that although it is found via minimizing the empirical risk $\mathcal{E}_n(f)$, it can eventually predict as well as the optimal predictor $\tilde{f}$, or it can give the best (the lowest risk) prediction for any unlabeled samples. In this sense, the generalization capability of a learning algorithm is a necessary requirement for any successful application, and furthermore, to characterize generalization capability of a learning algorithm requires in essence to decipher how close $\hat{f}$ is from $\tilde{f}$. This is a very difficult issue in general [4]. In statistical learning framework, we usually consider how close the expected risk $\mathcal{E}(\hat{f})$ is from $\mathcal{E}(\tilde{f})$, or equivalently, how small can we expect the difference $\mathcal{E}(\hat{f}) - \mathcal{E}(\tilde{f})$ to be. Whenever $\mathcal{E}(\hat{f})$ eventually approaches to $\mathcal{E}(\tilde{f})$, we say that the learning algorithm is consistent. Our aim in this paper is to conclude the consistency of the ERM algorithm with uniformly ergodic Markov chain samples, and provide the generalization bound estimations for the ERM algorithm.

Since $\hat{f}$ is dependent on the sample set S, in other words, the minimization (1) is taken over the discrete quantity $\mathcal{E}_n(f)$, intuitively, we have to estimate the capacity of the function set $\mathcal{F}$. Here the capacity of the function set $\mathcal{F}$ is measured by the covering number in this paper.

Definition 2 ([4]). For a subset $\mathcal{F}$ of a metric space and $\varepsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \varepsilon)$ of the function set $\mathcal{F}$ is the minimal integer $b \in \mathbb{N}$ such that there exist b disks with radius ε covering $\mathcal{F}$.

To establish the generalization bounds of the ERM algorithm with uniformly ergodic Markov chain samples, we give some basic assumptions on the hypothesis space $\mathcal{F}$ and the loss function $\ell(f, z)$:

(i) We suppose that $\mathcal{F}$ is contained in a ball of a Hölder space $\mathcal{C}^p(X)$ on a compact subset of a Euclidean space $\mathbb{R}^d$ for some $p > 0$. Here the Hölder space $\mathcal{C}^p(X)$ is defined as the space of all continuous functions on X with the following norm finite [11]:

$$\|f\|_{\mathcal{C}^p(X)} = \|f\|_\infty + |f|_{\mathcal{C}^p(X)}, \qquad |f|_{\mathcal{C}^p(X)} := \sup_{x_1 \neq x_2, x_1, x_2 \in X} \frac{|f(x_1) - f(x_2)|}{(d(x_1, x_2))^p},$$

where $d(\cdot, \cdot)$ is the metric defined on X.

(ii) Let

$$M \doteq \sup_{f \in \mathcal{F}} \max_{z \in \mathbf{Z}} \ell(f, z), \qquad L \doteq \sup_{g_1, g_2 \in \mathcal{F}, g_1 \neq g_2} \max_{z \in \mathbf{Z}} \frac{|\ell(g_1, z) - \ell(g_2, z)|}{\|g_1 - g_2\|_{\mathcal{C}^p(X)}}.$$

We suppose that M and L both are finite in this paper.

By the basic assumption (i), there exists a constant $C_0 > 0$ such that for any $\varepsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \varepsilon)$ of $\mathcal{F}$ in $\mathcal{C}^p(X)$ with the metric $\|\cdot\|_{\mathcal{C}^p(X)}$ satisfies (see [21]),

$$\mathcal{N}(\mathcal{F}, \varepsilon) \leq \exp\{C_0 \varepsilon^{\frac{-2d}{p}}\}. \tag{2}$$

The interested reader can consult [4] for examples of the hypothesis space $\mathcal{F}$.

For this end, we will first study the bound on the uniform convergence of the ERM algorithm with u.e.M.c. samples in the next section. In doing so, we will apply the following Hoeffding's inequality on u.e.M.c., which established by Glynn and Ormoneit in [20] and a useful lemma established by Cucker and Smale in [22].

Lemma 1 (*[20]*). *Suppose that $\{\mathbf{z}_i\}_{i \geq 0}$ is a uniformly ergodic Markov taking values in a state space Ω, $g : \Omega \to \mathbb{R}$ is a real function, $U_i = g(\mathbf{z}_i)$, and $S_n = \sum_{i=0}^{n-1} U_i$ for any integer n. If the norm of g is defined by*

$$\|g\|_\infty \doteq \sup\{|g(\mathbf{z})| : \mathbf{z} \in \Omega\},$$

and $\|g\|_\infty \leq \infty$, then for any $\varepsilon > 0$, and any n not less than $\frac{2m\|g\|_\infty}{\lambda\varepsilon}$, the inequality

$$\mathrm{Prob}_z\{S_n - \mathrm{E}[S_n] \geq n\varepsilon\} \leq \exp\left\{\frac{-(n\varepsilon\lambda - 2\|g\|_\infty m)^2}{2n\|g\|_\infty^2 m^2}\right\}$$

is valid, where $\mathrm{Prob}_z\{A\} = \mathrm{Prob}\{A|\mathbf{z}_0 = z\}$.

Lemma 2 (*[22]*). *Let $c_1,\ c_2 > 0$, and $s > q > 0$. Then the equation*

$$x^s - c_1 x^q - c_2 = 0$$

has a unique positive zero x^*. In addition

$$x^* \leq \max\{(2c_1)^{1/(s-q)}, (2c_2)^{(1/s)}\}.$$

3. Uniform convergence bound

In this section we quantitatively study uniform convergence of the ERM algorithm with u.e.M.c. samples. To be more precise, we provide an upper bound estimation on the rate of probabilistic convergence of the ERM algorithm with u.e.M.c. samples.

Theorem 1. *Let $\underline{Z}(\lambda, m)$ be a uniformly ergodic Markov chain and $\mathcal{N}(\mathcal{F}, \varepsilon)$ be the covering number of $\mathcal{F}$. Then for any $\varepsilon > 0$, and any $n \geq \frac{2Mm}{\lambda\varepsilon}$, there holds*

$$\mathrm{Prob}\left\{\sup_{f \in \mathcal{F}} |\mathcal{E}(f) - \mathcal{E}_n(f)| \geq \varepsilon\right\} \leq 2\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{4L}\right) \exp\left\{\frac{-(n\varepsilon\lambda - 4Mm)^2}{8nM^2m^2}\right\}. \tag{3}$$

Proof. For any $i \in \{0, 1, \ldots, n-1\}$, let $U_i = \mathrm{E}[\ell(f, \mathbf{z}_0)] - \ell(f, \mathbf{z}_i)$. Then $\mathcal{E}(f) - \mathcal{E}_n(f) = \frac{1}{n}\sum_{i=0}^{n-1} U_i$, and

$$|U_i| = |\ell(f, \mathbf{z}_i) - \mathrm{E}[\ell(f, \mathbf{z}_0)]| \leq M.$$

Replacing $\|g\|_\infty$ by M in Lemma 1, and by the fact that $\text{Prob}(A) = \text{E}[\text{Prob}_z(A)]$, we obtain

$$\text{Prob}\{|\mathcal{E}(f) - \mathcal{E}_n(f)| \geq \varepsilon\} \leq 2\exp\left\{\frac{-(n\varepsilon\lambda - 2Mm)^2}{2nM^2m^2}\right\}. \tag{4}$$

Set

$$L_S(f) = \mathcal{E}(f) - \mathcal{E}_n(f), \qquad k = \mathcal{N}(\mathcal{F}, \varepsilon),$$

and let $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \cdots \cup \mathcal{F}_k$. By the same argument conducted as that in [4], we have

$$\text{Prob}\left\{\sup_{f\in\mathcal{F}} |\mathcal{E}(f) - \mathcal{E}_n(f)| \geq \varepsilon\right\} \leq \sum_{j=1}^{k} \text{Prob}\left\{\sup_{f\in\mathcal{F}_j} |\mathcal{E}(f) - \mathcal{E}_n(f)| \geq \varepsilon\right\}. \tag{5}$$

To estimate the term on the right-hand side of the above inequality, we define balls $\mathcal{F}_j, j \in \{1, 2, \ldots, k\}$ to be a cover of $\mathcal{F}$ with center at f_j, and radius ε. Then, for all $f \in \mathcal{F}_j$,

$$\begin{aligned} |L_S(f) - L_S(f_j)| &\leq \text{E}[|\ell(f, z) - \ell(f_j, z)|] + \frac{1}{n}\sum_{i=0}^{n-1} |\ell(f, z_i) - \ell(f_j, z_i)| \\ &\leq 2L \cdot \|f - f_j\|_{\mathcal{C}^p(X)} \\ &\leq 2L\varepsilon. \end{aligned}$$

It follows that for all $f \in \mathcal{F}_j$

$$\sup_{f\in\mathcal{F}_j} |L_S(f)| \geq 4L\varepsilon \Longrightarrow |L_S(f_j)| \geq 2L\varepsilon.$$

By inequality (4), we thus conclude that for any $j \in \{1, 2, \ldots, k\}$,

$$\text{Prob}\left\{\sup_{f\in\mathcal{F}_j} |L_S(f)| \geq 4L\varepsilon\right\} \leq 2\exp\left\{\frac{-(2nL\varepsilon\lambda - 2mM)^2}{2nm^2M^2}\right\}. \tag{6}$$

Combining inequalities (5) and (6) and replacing ε by $\frac{\varepsilon}{4L}$ then gives Theorem 1. This completes the proof of Theorem 1. □

Remark 2. Theorem 1 shows that as long as the covering number of the hypothesis space $\mathcal{F}$ is finite, the empirical risk $\mathcal{E}_n(f)$ uniformly converges to the expected risk $\mathcal{E}(f)$, and the convergence speed may be exponential. This assertion is well known for the ERM algorithm with i.i.d. samples (see [1,2,4]). We have generalized this classical result to the u.e.M.c. samples. In addition, bound (3) have the same order with that obtained by Vapnik in [1], and by Cucker and Smale in [4].

Now we derive the generalization bounds and conclude the consistency of the ERM algorithm with u.e.M.c. samples by applying the uniform convergence bound estimation results obtained in Theorem 1.

Observing from Theorem 1 that whenever $\lambda \leq \varepsilon$, the exponential in (3) becomes

$$\frac{-(n\varepsilon\lambda - 4mM)^2}{8nM^2m^2} \leq \frac{-\varepsilon^2(\lambda^2 n - 8mM)}{8m^2M^2}.$$

By assumption (2), we have

$$\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{4L}\right) \leq \exp\left\{C_0\left(\frac{\varepsilon}{4L}\right)^{-\frac{2d}{p}}\right\}.$$

So, by Theorem 1, we have that for any $\varepsilon > 0$,

$$\text{P}\left\{\sup_{f\in\mathcal{F}} |\mathcal{E}(f) - \mathcal{E}_n(f)| \geq \varepsilon\right\} \leq 2\exp\left\{C_0\left(\frac{\varepsilon}{4L}\right)^{-\frac{2d}{p}} - \frac{\varepsilon^2(\lambda^2 n - 8mM)}{8m^2M^2}\right\}. \tag{7}$$

Now we rewrite inequality (7) in an equivalent form: For any $\delta \in (0, 1]$, let

$$\exp\left\{C_0\left(\frac{\varepsilon}{4L}\right)^{-\frac{2d}{p}} - \frac{\varepsilon^2(\lambda^2 n - 8mM)}{8m^2M^2}\right\} = \delta.$$

We have

$$\varepsilon^{(2+\frac{2d}{p})} - \frac{8\ln(1/\delta)m^2M^2}{(\lambda^2 n - 8mM)}\varepsilon^{\frac{2d}{p}} - \frac{8C_0(4L)^{\frac{2d}{p}}m^2M^2}{(\lambda^2 n - 8mM)} = 0.$$

By Lemma 2, we have that this equation with respect to ε has a unique positive zero ε^*, and

$$\varepsilon^* \le \varepsilon(n, \delta) \doteq \max\left\{\left[\frac{16\ln(1/\delta)m^2M^2}{\lambda^2 n - 8mM}\right]^{\frac{1}{2}}, \left[\frac{16m^2M^2C_0(4L)^{\frac{2d}{p}}}{\lambda^2 n - 8mM}\right]^{\frac{p}{2p+2d}}\right\}.$$

The solution $\varepsilon(n, \delta)$ is used to solve the inequality

$$\sup_{f\in\mathcal{F}}[\mathcal{E}(f) - \mathcal{E}_n(f)] \le \varepsilon(n, \delta).$$

As a result we obtain that with probability at least $1 - \delta$ for any function $f \in \mathcal{F}$, the inequality

$$\mathcal{E}(f) \le \mathcal{E}_n(f) + \varepsilon(n, \delta)$$

is valid. It is true as well for the function $\hat{f}$ that minimizing the empirical risk $\mathcal{E}_n(f)$ over $\mathcal{F}$. Thus the bound

$$\mathcal{E}(\hat{f}) \le \mathcal{E}_n(\hat{f}) + \varepsilon(n, \delta) \tag{8}$$

holds with probability at least $1 - \delta$.

On the other hand, for the same δ as above, we have that for the function $\tilde{f}$ minimizing the expected risk $\mathcal{E}(f)$ over $\mathcal{F}$, the following estimation

$$\mathcal{E}(\tilde{f}) \ge \mathcal{E}_n(\tilde{f}) - \varepsilon(n, \delta) \tag{9}$$

holds with probability $1 - \delta$.

Note that $\mathcal{E}_n(\tilde{f}) \ge \mathcal{E}_n(\hat{f})$. From inequalities (8) and (9), we thus deduce that with probability at least $1 - 2\delta$, the estimation

$$\mathcal{E}(\hat{f}) - \mathcal{E}(\tilde{f}) \le 2\varepsilon(n, \delta) \tag{10}$$

holds true.

Remark 3. (i) Bounds (8) and (10) show that the ERM algorithm are generalizable when applied to the u.e.M.c. samples. This assertion generalizes the previous results in [1] of i.i.d. samples to the u.e.M.c. samples.

(ii) By bound (10), we have

$$\mathcal{E}(\hat{f}) - \mathcal{E}(\tilde{f}) \to 0, \quad \text{as } n \to \infty.$$

This then shows that the ERM algorithm with u.e.M.c. samples is consistent. This conclusion extends the classical results in [1] to the case where the i.i.d. samples replaced by the u.e.M.c. samples. The difference is that there we use a simpler concept of capacity than in the classical model in [1].

4. Relative uniform convergence bound

In this section, our aim is to generalize the uniform convergence bounds established in the last section to the relative uniform convergence case. That is, we estimate the quantity (for any $\varepsilon > 0$)

$$\text{Prob}\left\{\sup_{f\in\mathcal{F}}\frac{|\mathcal{E}(f)-\mathcal{E}_n(f)|}{\sqrt{\mathcal{E}(f)}}\geq\varepsilon\right\} \tag{11}$$

by following the enlightening idea of [13,20]. Such a study is motivated by the observation that a uniform convergence bound fails to capture the phenomenon that for those functions $f\in\mathcal{F}$ for which the expected risk $\mathcal{E}(f)$ is small, the deviation $\mathcal{E}(f)-\mathcal{E}_n(f)$ is also small with large probability. However, the relative deviation $[\mathcal{E}(f)-\mathcal{E}_n(f)]/\sqrt{\mathcal{E}(f)}$ is more appropriate for capturing the phenomenon [1]. To develop an upper bound of term (11), we clearly need to assume that the denominator in term (11) does not take value 0. Accordingly, replacing $\mathcal{F}$ in the last section, we confine the function f to belong to a more restricted function set $\mathcal{H}$ in this section, where

$$\mathcal{H}=\{f\in\mathcal{F}\ :a\leq\|f\|_{\mathcal{C}^p(X)}\leq b\}$$

with two positive real number a and b (not necessary bounded).

Just as the role of Hoeffding's inequality played in the proof of Theorem 1, Markov's inequality will play a crucial role in the proof of the following estimation.

Theorem 2. *Let $\underline{Z}(\lambda,m)$ be a uniformly ergodic Markov chain. Then for any $\varepsilon>0$, any $n\geq\frac{2Mm}{\lambda\varepsilon}$, and for all $f\in\mathcal{H}$, the following inequality*

$$\text{Prob}\left\{\frac{\mathcal{E}_n(f)-\mathcal{E}(f)}{\sqrt{\mathcal{E}(f)}}\geq\varepsilon\right\}\leq\exp\left\{\frac{-(n\varepsilon\lambda\sqrt{aL}-2Mm)^2}{2nM^2m^2}\right\}$$

is valid.

Proof. For simplicity, we set $\mu=\mathcal{E}(f)$. Let $f_c(\mathbf{z})\doteq\ell(f,\mathbf{z})-\mu$, and set

$$S_n-n\mu\doteq\sum_{i=0}^{n-1}\ell(f,\mathbf{z}_i)-n\mu=\sum_{i=0}^{n-1}f_c(\mathbf{z}_i).$$

We then easily find that $\mathcal{E}_n(f)-\mathcal{E}(f)=\frac{1}{n}(S_n-n\mu)$. Under Assumption 1, it is known that (see [23])

$$|\mathrm{E}_z[f_c(\mathbf{z}_n)]|\leq M\cdot(1-\lambda)^{\lfloor n/m\rfloor},$$

where $\mathrm{E}_z(\cdot)\doteq\mathrm{E}[\cdot|\mathbf{z}_0=z]$. Hence $g(\mathbf{z})\doteq\sum_{n=1}^{\infty}\mathrm{E}_z[f_c(\mathbf{z}_n)]$ converges absolutely and

$$\|g\|_\infty\leq M\cdot m/\lambda, \tag{12}$$

where $\|g\|_\infty\doteq\sup\{|g(\mathbf{z})|:\mathbf{z}\in\mathbf{Z}\}$. Furthermore, g solves Poisson's equation

$$g(\mathbf{z})-\mathrm{E}_z[g(\mathbf{z}_1)]=f_c(\mathbf{z})$$

for any $\mathbf{z}\in\mathbf{Z}$. Observe that $D_i=g(\mathbf{z}_i)-\mathrm{E}_z[g(\mathbf{z}_i)|\mathbf{z}_0,\ldots,\mathbf{z}_{n-1}]$ is a martingale difference for $i\geq0$. Furthermore, it follows that

$$S_n-n\mu=\sum_{i=1}^{n}D_i+g(\mathbf{z}_0)-g(\mathbf{z}_n).$$

Then we conclude that for any $\theta>0$

$$\mathrm{E}_z\left[\exp(\theta(S_n-n\mu))\right]\leq\exp(2\theta\|g\|_\infty)\cdot\mathrm{E}_z\left[\exp\left(\theta\sum_{i=1}^{n}D_i\right)\right].$$

But

$$\mathrm{E}_z\left[\exp\left(\theta\sum_{i=0}^{n-1}D_i\right)\right]=\mathrm{E}_z\left[\exp\left(\theta\sum_{i=1}^{n}D_i\right)\right]\mathrm{E}_z[\exp(\theta D_n)|\mathbf{z}_0,\ldots,\mathbf{z}_{n-1}].$$

As D_i lies a.s. in an interval of length $2\|g\|_\infty$, we then have (see Lemma 8.1 in [24])

$$\mathrm{E}_z[\exp(\theta D_n)|\mathbf{z}_0,\ldots,\mathbf{z}_{n-1}]\leq\exp(\theta^2\|g\|_\infty^2/2).$$

Consequently, we get the inequality $\mathrm{E}_z\left[\exp(\theta(S_n - n\mu))\right] \le \exp(2\theta\|g\|_\infty + n\theta^2\|g\|_\infty^2/2)$. Taking expectation of both sides with respect to z, we thus obtain

$$\mathrm{E}\left[\exp(\theta(S_n - n\mu))\right] \le \exp(2\theta\|g\|_\infty + n\theta^2\|g\|_\infty^2/2).$$

By using Markov's inequality, we then deduce that for any $\theta > 0$,

$$\begin{aligned}\mathrm{Prob}\{S_n - n\mu \ge n\delta\mu\} &= \mathrm{Prob}\left\{e^{\theta(S_n - n\mu)} \ge e^{\theta n\delta\mu}\right\}\\ &\le \frac{\mathrm{E}[e^{\theta(S_n - n\mu)}]}{e^{\theta n\delta\mu}}\\ &\le \exp(2\theta\|g\|_\infty + n\theta^2\|g\|_\infty^2/2 - \theta n\delta\mu). \qquad (13)\end{aligned}$$

Taking $\theta = \frac{n\delta\mu - 2\|g\|_\infty}{n\|g\|_\infty^2}$, substituting θ into inequality (13) and using the bound (12), we obtain

$$\mathrm{Prob}\{S_n - n\mu \ge n\delta\mu\} \le \exp\left\{\frac{-(n\delta\lambda\mu - 2Mm)^2}{2nM^2m^2}\right\}.$$

Replacing δ by $\frac{\varepsilon}{\sqrt{\mu}}$, we have

$$\mathrm{Prob}\left\{\frac{\mathcal{E}_n(f) - \mathcal{E}(f)}{\sqrt{\mathcal{E}(f)}} \ge \varepsilon\right\} \le \exp\left\{\frac{-(n\varepsilon\lambda\sqrt{\mu} - 2mM)^2}{2nM^2m^2}\right\}.$$

Since for any $f \in \mathcal{H}$, we have $a \le \|f\|_{\mathcal{C}^p(X)} \le b$. By the basic assumption (ii), we then have $aL \le \mathcal{E}(f) \le bL$. Thus we conclude that for any $\varepsilon > 0$

$$\mathrm{Prob}\left\{\frac{\mathcal{E}_n(f) - \mathcal{E}(f)}{\sqrt{\mathcal{E}(f)}} \ge \varepsilon\right\} \le \exp\left\{\frac{-(n\varepsilon\lambda\sqrt{aL} - 2mM)^2}{2nM^2m^2}\right\}.$$

This completes the proof of Theorem 2. □

Theorem 3. *Let $\underline{Z}(\lambda, m)$ be a uniformly ergodic Markov chain and $\mathcal{N}(\mathcal{H}, \varepsilon)$ be the covering number of $\mathcal{H}$. Then for any $\frac{2}{3}(aL) > \varepsilon > 0$, and any $n \ge \frac{2Mm}{\lambda\varepsilon}$, the inequality*

$$\mathrm{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{|\mathcal{E}_n(f) - \mathcal{E}(f)|}{\sqrt{\mathcal{E}(f)}} \ge \varepsilon\right\} \le 2\mathcal{N}(\mathcal{H}, \varphi\varepsilon)\exp\left\{\frac{-(n\lambda\tau - 2Mm)^2}{2nM^2m^2}\right\}$$

holds, where $\tau = \frac{7a}{3\sqrt{bL}(b+4a)}\varepsilon$, and $\varphi = \frac{\sqrt{a}}{(b+4a)L^{\frac{3}{2}}}$.

Proof. We decompose the proof into two steps.

Step 1. Let $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots \cup \mathcal{H}_l$, then for any $\varepsilon > 0$, we have

$$\mathrm{Prob}\left\{\sup_{f\in\mathcal{H}} \frac{\mathcal{E}_n(f) - \mathcal{E}(f)}{\sqrt{\mathcal{E}(f)}} \ge \varepsilon\right\} \le \sum_{j=1}^{l} \mathrm{Prob}\left\{\sup_{f\in\mathcal{H}_j} \frac{\mathcal{E}_n(f) - \mathcal{E}(f)}{\sqrt{\mathcal{E}(f)}} \ge \varepsilon\right\}. \qquad (14)$$

Define

$$\phi(f) = (1-\delta)\mathcal{E}(f) - \mathcal{E}_n(f),$$

and $l = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{L}\right)$. Let the balls $\mathcal{H}_j$, $j \in \{1, 2, \ldots, l\}$ be a cover of $\mathcal{H}$ with center at f_j and radius ε/L. For any S and all $f \in \mathcal{H}_j$,

$$\begin{aligned}\phi(f) - \phi(f_j) &= (1-\delta)\mathcal{E}(f) - \mathcal{E}_n(f) - [(1-\delta)\mathcal{E}(f_j) - \mathcal{E}_n(f_j)]\\ &= [\mathcal{E}_n(f_j) - \mathcal{E}_n(f)] + (1-\delta)[\mathcal{E}(f) - \mathcal{E}(f_j)]\\ &\le L\cdot\|f_j - f\|_{\mathcal{C}^p(X)} + (1-\delta)L\cdot\|f_j - f\|_{\mathcal{C}^p(X)}\\ &\le \varepsilon(2-\delta).\end{aligned}$$

Since this holds for all S and all $f \in \mathcal{H}_j$, we find

$$\sup_{f\in\mathcal{H}_j} \phi(f) \geq 2\varepsilon(2-\delta) \Longrightarrow \phi(f_j) \geq \varepsilon(2-\delta).$$

Thus we obtain that for $j = 1, 2, \ldots, l$,

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}_j} \phi(f) \geq 2\varepsilon(2-\delta)\right\} \leq \text{Prob}\left\{\phi(f_j) \geq \varepsilon(2-\delta)\right\}. \tag{15}$$

Step 2. We denote by I_1 the quantity on the right-hand side of inequality (15), and by I_2 the quantity on the left-hand side of inequality (15). Then, take $\delta = \frac{\varepsilon}{\mathcal{E}(f_j)}$, suppose that $0 < \varepsilon \leq \frac{2}{3}(aL)$, and use the similar method of [13], we have

$$\begin{aligned}
I_1 &= \text{Prob}\left\{\phi(f_j) \geq \varepsilon(2-\delta)\right\} \\
&= \text{Prob}\left\{\mathcal{E}(f_j) - \mathcal{E}_n(f_j) \geq \delta\mathcal{E}(f_j) + \varepsilon(2-\delta)\right\} \\
&= \text{Prob}\left\{\mathcal{E}(f_j) - \mathcal{E}_n(f_j) \geq \varepsilon + \varepsilon\left(2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right)\right\} \\
&= \text{Prob}\left\{\frac{\mathcal{E}_n(f_j) - \mathcal{E}(f_j)}{\sqrt{\mathcal{E}(f_j)}} \geq \frac{\varepsilon}{\sqrt{\mathcal{E}(f_j)}} + \frac{\varepsilon}{\sqrt{\mathcal{E}(f_j)}}\left(2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right)\right\} \\
&\leq \text{Prob}\left\{\frac{\mathcal{E}(f_j) - \mathcal{E}_n(f_j)}{\sqrt{\mathcal{E}(f_j)}} \geq \frac{\varepsilon}{\sqrt{bL}}\left(3 - \frac{\varepsilon}{aL}\right)\right\} \\
&\leq \text{Prob}\left\{\frac{\mathcal{E}(f_j) - \mathcal{E}_n(f_j)}{\sqrt{\mathcal{E}(f_j)}} \geq \frac{7\varepsilon}{3\sqrt{bL}}\right\} \\
I_2 &= \text{Prob}\left\{\sup_{f\in\mathcal{H}_j} \phi(f) \geq 2\varepsilon(2-\delta)\right\} \\
&= \text{Prob}\left\{\sup_{f\in\mathcal{H}_j}\left[\mathcal{E}(f) - \mathcal{E}_n(f) - \frac{\varepsilon\mathcal{E}(f)}{\mathcal{E}(f_j)}\right] \geq 2\varepsilon\left[2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right]\right\} \\
&\geq \text{Prob}\left\{\sqrt{aL}\sup_{f\in\mathcal{H}_j}\left[\frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} - \frac{bL\varepsilon}{\sqrt{aL}\mathcal{E}(f_j)}\right] \geq 2\varepsilon\left[2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right]\right\} \\
&\geq \text{Prob}\left\{\sup_{f\in\mathcal{H}_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \frac{bL\varepsilon}{\sqrt{aL}\mathcal{E}(f_j)} + \frac{2\varepsilon}{\sqrt{aL}}\left[2 - \frac{\varepsilon}{\mathcal{E}(f_j)}\right]\right\} \\
&\geq \text{Prob}\left\{\sup_{f\in\mathcal{H}_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon\left[\frac{b}{a\sqrt{aL}} + \frac{2}{\sqrt{aL}}\left(2 - \frac{\varepsilon}{bL}\right)\right]\right\} \\
&\geq \text{Prob}\left\{\sup_{f\in\mathcal{H}_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \frac{(b+4a)\sqrt{L}}{\sqrt{a}}\varepsilon\right\}.
\end{aligned}$$

Then, from inequality (15), we obtain

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}_j} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \geq \varepsilon'\right\} \leq \text{Prob}\left\{\frac{\mathcal{E}(f_j) - \mathcal{E}_n(f_j)}{\sqrt{\mathcal{E}(f_j)}} \geq \tau\right\}$$

with

$$\varepsilon' = \frac{(b+4a)\sqrt{L}}{\sqrt{a}}\varepsilon, \qquad \tau = \frac{7}{3\sqrt{bL}}\varepsilon.$$

By inequality (14) and Theorem 2, this implies

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}}\frac{\mathcal{E}(f)-\mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}}\geq\varepsilon'\right\}\leq\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{L}\right)\exp\left\{\frac{-(n\tau\lambda\sqrt{aL}-2Mm)^2}{2nM^2m^2}\right\}.$$

Similarly we can justify

$$\text{Prob}\left\{\sup_{f\in\mathcal{H}}\frac{\mathcal{E}_n(f)-\mathcal{E}(f)}{\sqrt{\mathcal{E}(f)}}\geq\varepsilon'\right\}\leq\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{L}\right)\exp\left\{\frac{-(n\tau\lambda\sqrt{aL}-2Mm)^2}{2nM^2m^2}\right\}.$$

Combining these two inequalities, and replacing ε by $\frac{\sqrt{a}}{(b+4a)\sqrt{L}}\varepsilon$, we can complete the proof of Theorem 3. □

Remark 4. (i) Theorem 3 shows that as long as the covering number of the function set $\mathcal{H}$ is finite, the empirical risk $\mathcal{E}_n(f)$ uniformly converges to the expected risk $\mathcal{E}(f)$, and the confidence interval for the expected risk based on Theorem 3 is smaller than that based on Theorem 1 (this is the reason why Vapnik [1], Bousquet [2] bounded the term (11)).

(ii) Comparing the uniform convergence bound in Theorem 1 and the relative uniform convergence bound in Theorem 3 with these bounds based on mixing samples (e.g. α-mixing and β-mixing) in [10,12], we can find that these bounds for mixing samples have the rate $O(\exp(-n^{(\alpha)}))$, where n is the number of samples and $n^{(\alpha)}<n$ is the "effective number of observations". Therefore, the bounds for mixing samples have worse rate than that for i.i.d. samples and u.e.M.c. samples. This implies that mixing samples (e.g. α-mixing and β-mixing) contain less information than i.i.d. samples and u.e.M.c. samples.

Now we begin to establish the generalization bounds by using the relative uniform convergence bound of the ERM algorithm with u.e.M.c. samples.

Similarly, we suppose that $\lambda\leq\tau$, the exponential of Theorem 3 becomes

$$\frac{-(n\lambda\tau-2mM)^2}{2nM^2m^2}\leq\frac{-\tau^2(\lambda^2n-4mM)}{2m^2M^2}.$$

By assumption (2), we have

$$\mathcal{N}\left(\mathcal{H},\frac{\sqrt{a}\varepsilon}{(b+4a)L^{\frac{3}{2}}}\right)\leq\exp\left\{C_0\left(\frac{\sqrt{a}\varepsilon}{(b+4a)L^{\frac{3}{2}}}\right)^{-\frac{2d}{p}}\right\}.$$

By Theorem 3, we have that for any $\varepsilon>0$,

$$\text{P}\left\{\sup_{f\in\mathcal{H}}\frac{|\mathcal{E}(f)-\mathcal{E}_n(f)|}{\sqrt{\mathcal{E}(f)}}\geq\varepsilon\right\}\leq 2\exp\left\{C_0\,(\varphi\varepsilon)^{-\frac{2d}{p}}-\frac{\tau^2(\lambda^2n-4mM)}{2m^2M^2}\right\},\tag{16}$$

where $\varphi=\frac{\sqrt{a}}{(b+4a)L^{\frac{3}{2}}}$. Now we rewrite inequality (16) in an equivalent form: For any $\delta\in(0,1]$, let

$$\exp\left\{C_0\left(\frac{\sqrt{a}\varepsilon}{(b+4a)L^{\frac{3}{2}}}\right)^{-\frac{2d}{p}}-\frac{\tau^2(\lambda^2n-4mM)}{2m^2M^2}\right\}=\delta.$$

We have

$$\varepsilon^{(2+\frac{2d}{p})}-\frac{18\ln(1/\delta)m^2M^2b^2}{49a^2(\lambda^2n-4mM)}\varepsilon^{\frac{2d}{p}}-\frac{18C_0(L)^{\frac{3d+2p}{p}}m^2M^2b^2(b+4a)^{\frac{2p+2d}{p}}}{49a^{\frac{2p+d}{p}}(\lambda^2n-4mM)}=0.$$

By Lemma 2, we have that this equation with respect to ε has a unique positive zero ε^*, and

$$\varepsilon^* \le \varepsilon^*(n,\delta) \doteq \max\left\{\frac{6mMb}{7a}\left(\frac{\ln(1/\delta)}{\lambda^2 n - 4mM}\right)^{\frac{1}{2}}, \omega\left[\frac{18m^2M^2C_0b^2L^{\frac{d}{p}}}{49a^{\frac{2p+d}{p}}(\lambda^2 n - 4mM)}\right]^{\frac{p}{2p+2d}}\right\},$$

where $\omega = (b+4a)L$. The solution $\varepsilon^*(n,\delta)$ is used to solve the inequality

$$\sup_{f\in\mathcal{H}} \frac{\mathcal{E}(f) - \mathcal{E}_n(f)}{\sqrt{\mathcal{E}(f)}} \le \varepsilon^*(n,\delta).$$

As a result we obtain that with probability at least $1-\delta$ for the function $\bar{f}$ that minimizing the empirical risk $\mathcal{E}_n(f)$ over $\mathcal{H}$, the bound

$$\mathcal{E}(\bar{f}) \le \mathcal{E}_n(\bar{f}) + \frac{\varepsilon^*(n,\delta)}{2}\left(1 + \sqrt{1 + \frac{\mathcal{E}_n(\bar{f})}{\varepsilon^*(n,\delta)}}\right) \tag{17}$$

holds. By the similar argument with inequalities (9), we have that for the same δ as above, and for the function f' minimizing the expected risk $\mathcal{E}(f)$ over $\mathcal{H}$, the inequality

$$\mathcal{E}(f') \ge \mathcal{E}_n(f') - \varepsilon(n,\delta) \tag{18}$$

holds with probability $1-\delta$.

By inequalities (17) and (18), we thus deduce that with probability at least $1-2\delta$, the estimation

$$\mathcal{E}(\bar{f}) - \mathcal{E}(f') \le \varepsilon(n,\delta) + \frac{\varepsilon^*(n,\delta)}{2}\left(1 + \sqrt{1 + \frac{\mathcal{E}_n(\bar{f})}{\varepsilon^*(n,\delta)}}\right) \tag{19}$$

is valid.

Remark 5. Bounds (17) and (19) describe the generalization performance of the ERM algorithm with u.e.M.c. observations in the given function set $\mathcal{H}$: Bound (17) evaluates the risk for the chosen function in the target function set $\mathcal{H}$, and bound (19) evaluates how close this risk is to the smallest possible risk for the target functions set $\mathcal{H}$.

5. Conclusion

In this paper, we have studied the extension problem of statistical learning theory (SLT) from the classical independent and identically distributed (i.i.d.) sampling to the uniformly ergodic Markov chain (u.e.M.c.) sampling. Like i.i.d. sampling, the u.e.M.c. sampling is a naturally and extensively appeared random sampling mechanism, especially in the study of time or content-based pattern recognition or biological sequence analysis. The fundamental problems in SLT are evaluation of generalization performance and consistency of the ERM algorithm. We have extended the classical generalization bound estimations of the ERM algorithm through establishing a series of new bounds on the rate of uniform convergence and relative uniform convergence. From the established generalization bound estimations, we have draw a conclusion that the ERM algorithm with u.e.M.c. samples is consistent. The obtained results perfectly extended the well-known statistical learning theory for the ERM algorithm with i.i.d. observations in [1]. To our knowledge, the results here are the first explicit bounds on the rate of convergence on this topic.

There have been several other attempts to extend the classical SLT from i.i.d. samples to dependent observations (see e.g. [7,9,12]). Different from those works, the uniform convergence bound and the relative uniform convergence bound obtained for the ERM algorithm with u.e.M.c samples have the same convergence order with that obtained by Vapnik in [1], by Cucker and Smale in [4] for i.i.d. samples. It is worth noting that among the existing attempts to generalize the learning theory from i.i.d. samples to dependent samples, there is still no convergence bound exactly preserving the same order with that for i.i.d. observations (say, those for α-mixing sequence [10]; those for exponentially

strongly mixing sequence [9,12]; those for β-mixing sequence [10]). In particular, under the same conditions (e.g. hypothesis space), we can find that the generalization bounds obtained in this paper also have the same order with that obtained for i.i.d. observations. In addition, the generalization bounds established in this paper are based on the assumption that the u.e.M.c. $\underline{Z}(\lambda, m)$ is stationary. In fact, the u.e.M.c. $\underline{Z}(\lambda, m)$ may not stationary, in this case we can also study the performance bounds by using the similar method of [11].

Along the line of the present work, several open problems deserve further research. For example, how to control the generalization ability of the ERM algorithm with u.e.M.c. samples? What is the essential difference of generalization ability of the ERM algorithm with i.i.d. samples and u.e.M.c. samples? how to develop a lower bound estimation on the generalization ability of the ERM algorithm with u.e.M.c. samples? All these problems are under our current investigation.

Acknowledgments

The authors would like to thank Xiangyu Chang for interesting discussions. We would like to thank the referees for their careful reading and helpful comments on the paper.

References

[1] V. Vapnik, Statistical Learning Theory, John Wiley, New York, 1998.
[2] O. Bousquet, New approaches to statistical learning theory, Ann. Inst. Statist. Math. 55 (2003) 371–389.
[3] D.R. Chen, Q. Wu, Y.M. Ying, D.X. Zhou, Support vector machine soft margin classifiers: Error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.
[4] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2002) 1–49.
[5] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, Cambridge, 2007.
[6] S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41 (2004) 279–305.
[7] I. Steinwart, D. Hush, C. Scovel, Learning from dependent observations, J. Multivariate Anal. 100 (1) (2009) 175–194.
[8] B. Yu, Rates of convergence for empirical processes of stationary mixing sequences, Ann. Probab. 22 (1994) 94–114.
[9] S. Modha, E. Masry, Minimum complexity regression estimation with weakly dependent observations, IEEE Trans. Inform. Theory 42 (1996) 2133–2145.
[10] M. Vidyasagar, Learning and Generalization with Applications to Neural Networks, 2nd edition, Springer, London, 2002.
[11] S. Smale, D.X. Zhou, Online learning with Markov sampling, http://www6.cityu.edu.hk/ma/people/dxzhou/SmaleZhou0708.pdf, 2008.
[12] B. Zou, L.Q. Li, The performance bounds of learning machines based on exponentially strongly mixing sequence, Comput. Math. Appl. 53 (7) (2007) 1050–1058.
[13] Bin Zou, Luoqing Li, Zongben Xu, The generalization performance of ERM algorithm with strongly mixing observations, Mach. Learn. (2009), doi:10.1007/s10994-009-5104-z.
[14] C. Andrieu, N.de. Freitas, A. Doucet, M. Jordan, An introduction to MCMC learning, Mach. Learn. 50 (2003) 5–43.
[15] W.R. Gilks, P. Wild, Adaptive rejection sampling for Gibbs sampling, Appl. Statist. 51 (2) (1992) 337–348.
[16] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic Stability, Springer-Verlag, 1993.
[17] W.K. Hastings, Monte Carlo sampling methods using Markov chain and their applications, Biometrika 57 (1970) 97–109.
[18] G.L. Jones, On the Markov chain central limit theorem, Probab. Surveys 1 (2004) 299–320.
[19] D. Aldous, L. Lovász, P. Winkler, Mixing times for uniformly ergodic Markov chains, Stochastic Process. Appl. 71 (1997) 165–185.
[20] P.W. Glynn, D. Ormoneit, Hoeffding's inequality for uniformly ergodic Markov chains, Statist. Probab. Lett. 56 (2002) 143–146.
[21] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1743–1752.
[22] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, Found. Comput. Math. 2 (2002) 413–428.
[23] S. Asmussen, P.W. Glynn, H. Thorisson, Stationarity detection in the initial transient problem, ACM Trans. Model. Comput. Simul. 2 (1992) 130–157.
[24] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1996.